What issues are raised by evaluating problem-based undergraduate medical curricula? Making healthy connections across the literature

Gillian Maudsley MB ChB, FRCPath, MPH (dist), MFPHM, MEd (dist)
Department of Public Health, The University of Liverpool, Liverpool, UK

Abstract
There have been various waves and ripples of undergraduate medical curricular reform over recent decades. Such programme-wide innovation can attract exceptional suspicion and scrutiny about its worth, fuelled by adverse reactions to change. It also raises expectations of what programme evaluation can reveal about the experience, achievements and ‘added value’ of the medical students or doctors. Problem-based undergraduate medical curricula are expected, for example, to fulfil elusive ‘true’ (outcome) descriptors of educational quality. The aim here was to explore the rationale, approach and challenges for undergraduate medical programme evaluation, particularly for problem-based curricula. The main focus was on internal, formative evaluation: Why undertake educational evaluation? What is the evaluative context of problem-based learning? What philosophy has guided programme evaluation of innovative undergraduate medical curricula? What can educational, health care and other evaluation frameworks offer? What can be learned from examples of published accounts of problem-based undergraduate medical programme evaluation? Where to from here? Ideally, evaluators of medical education should be explicit about their rationale, highlight local curricular context and special features, balance process-measures with outcome-measures (including unplanned outcomes), and be eclectic in methods. ‘Healthier links’ with the health care evaluation and educational evaluation literature could be beneficial. It remains to be seen, however, whether the medical educational changes that the General Medical Council’s Tomorrow’s Doctors triggered in the United Kingdom will stimulate important advances in educational evaluation.

Introduction
There have been various waves and ripples of undergraduate medical curricular reform over recent decades. Such programme-wide innovation can attract exceptional suspicion and scrutiny about its worth, fuelled by adverse reactions to change. It also raises expectations of what programme evaluation can reveal about the experience, achievements and ‘added value’ of the medical students or doctors. (Here programme is used to mean the overall course of study leading to the university award of a medical degree.)

Acknowledging that ‘innovation’ is a relative concept, Friedman et al. (1990) defined innovative undergraduate medical curricula (or component ‘tracks’ for particular cohorts) as:

- being wholly (or mostly) innovative;
• having an explicit sense of purpose (as with problem-based and/or community-orientated programmes);
• providing a clearly non-traditional educational experience for all the medical students, or a distinct subgroup following a different ‘track’.

For North America, reporting in 1990, Friedman et al. identified 10 innovative curricula (e.g. McMaster, New Mexico, and Sherbrooke), and Vernon (1995) identified 14 uniformly problem-based curricula (including ‘tracks’) in 1992–93. Elsewhere, for example, the established problem-based curricula at Newcastle (Australia) and Maastricht (Holland) provided innovative prototypes. Several British medical schools, such as Liverpool, Manchester (O’Neill et al. 1999) and Glasgow, underwent large-scale problem-based transformation in the mid-1990s.

Critics expect many problem-based undergraduate medical curricula continually to justify their existence (rather than show their special effects), often invoking traditional curricula as ‘gold standards’ of unquestionable authenticity. Stretching the problem-based learning (PBL) label around any active learning that uses a ‘problem’ is unhelpful. Such ‘conceptual fog’ then complicates comparisons between problem-based and traditional curricula (Maudsley 1999a).

The educational environment of British medical schools is becoming more evaluative. Firstly, the last General Medical Council (GMC) recommendations for undergraduate education (Tomorrow’s Doctors) promoted less factual overload, and more student-centred educational renewal, good practice and problem-orientation, to be evaluated under ‘continuing review’ (GMC 1993, p. 24). The GMC’s renewed vocational focus built on global aims for graduates to meet changing population health and health care needs (World Federation for Medical Education 1988; Walton 1994). Secondly, the National Health Service (NHS) is a partner in undergraduate medical education, the main source of work for its graduates, and the honorary clinical contract-holder for clinical academics; how it evaluates its own activities may well influence medical schools. Developments over the last decade include: corporate responsibility for continuously improving service quality (clinical governance, clinical revalidation); evidence-based health care; and strengthening the public health function (Donaldson 2001), which includes health care surveillance and evaluation activities. Thirdly, doctors’ increasing obligations for continuing professional development (Brigley et al. 1997) include their role as educators (GMC 1999). Fourthly, university factors include the Dearing review of higher education (National Committee of Inquiry into Higher Education 1997) and the Quality Assurance Agency for Higher Education (1997), evaluating students’ experience formally in broad subject-areas to improve standards, i.e. ‘subject review’.

The medical education literature tends to focus more on what educational evaluation finds than the rationale for the approach taken. The aim here was to explore the rationale, approaches and challenges for undergraduate medical programme evaluation, particularly for problem-based curricula. The main focus here was formative (Nevo 1986) (‘for feedback’, improvement, development) evaluation rather than the obligatory ‘summative’ (i.e. that ‘counts’ for accreditation) type:
• Why undertake educational evaluation?
• What is the evaluative context of problem-based learning?
• What philosophy has guided programme evaluation of innovative undergraduate medical curricula? Examples.
• What can educational and health care evaluation frameworks offer?
• What can be learned from examples of published accounts of problem-based undergraduate medical programme evaluation?
• Where to from here?

Searching for key publications involved several electronic databases:
• ‘ERIC’ 1966-June 2000.
• Science Citation Index Expanded and Social Sciences Citation Index on ‘ISI (Institute for Scientific Information) Web of Science’ 1981-March 2001.

Key search-terms comprised variations on: course/curricular/programme/outcome evaluation, compa-
rative analysis, evaluation methods against (problem-based) (innovative) undergraduate medical education/medical schools/medical students. Further publications were identified from this literature.

**Why undertake educational evaluation?**

Every adult education program is unique psychosocial drama... and the publicly stated reasons for conducting an evaluation may bear little relation to the private agendas possessed by the actors in the drama.

(Brookfield 1986 p. 266)

The term educational evaluation includes many diverse activities (Nevo 1986) that by judging merit can become very political. They may not even have formative or summative purpose, serving other purposes, e.g. psychosocial or administrative. Cantillon (1999) suggested that if more published evaluations of medical education appeared, this might attract much-needed institutional and financial support for innovation.

Developing on from Stufflebeam’s work, Nevo (1986) outlined 10 dimensions along which educational evaluation can differ (Box 1), i.e. by its: definition; function; focus/unit of study; information requirements; criteria for judging worth/value; ‘audience’; process; methods; evaluator’s identity; and standards.

Appraisal of evaluation evidence is therefore compromised when published reports in the medical education literature omit many of the clues about these dimensions.

Here, educational evaluation will refer to a whole curriculum/programme and broadly mean: a systematic process that judges the worth of an educational programme via quantitative and/or qualitative data analysis, consistent with the evaluation question, and aims to improve students’ experience and achievements. This may or may not focus on achievement of objectives/outcomes or providing information for decision-makers (Alkin 1975; Nevo 1986).

Of Nevo’s 10 dimensions, the function, i.e. specific purpose (Why evaluate at all?), of the evaluation is fundamental, although the stated and actual answer might differ greatly. The answer could be to stimulate or maintain curricular renewal. Alkin (1975) identified a crucial ‘Who cares?’ question. Establishing to whom the evaluation results matter (or maybe, unfortunately, ‘Who is “shouting loudest” for evidence?’) relates to Nevo’s (1986) dimensions of function and intended audience.

By the nature of reaction to change, innovative undergraduate medical education potentially has many concerned detractors to convince, placate, or even at least prevent from undermining staff and student efforts. Demands for ‘hard’ empirical outcome-evidence – Does a ‘new curriculum’ work (effective)? Does the product justify the effort (efficient)? – can dominate such evaluation, and useful outcome measures are elusive.

The ‘Why evaluate?’ question might appear simplistic until trying to answer it. Friedman et al. (1990) concluded that, by the early 1990s, the function of ‘outcome-driven’ comparative evaluations of innovative undergraduate medical curricula had changed. They noted that legitimating each new curriculum had given way to exploring distinctiveness, e.g. longer-term, unique effects, because competent graduates were emerging to provide some reassurance. While traditional curricula improve and incorporate good educational practice incrementally, arguably, marked innovative transformation still requires local legitimacy. Comprehensive problem-based curricula usually differ sufficiently from traditional counterparts in overall philosophy, student experience and local acceptance to provide a challenging context for evaluation.

<table>
<thead>
<tr>
<th>Box 1 Nevo’s 10 dimensions of educational evaluation. Source: after Nevo (1986)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How is evaluation defined?</strong></td>
</tr>
<tr>
<td><strong>What are the functions of evaluation?</strong></td>
</tr>
<tr>
<td><strong>What are the objects of evaluation?</strong></td>
</tr>
<tr>
<td><strong>What kinds of information should be collected regarding each object?</strong></td>
</tr>
<tr>
<td><strong>What criteria should be used to judge the merit and worth of an evaluation-object?</strong></td>
</tr>
<tr>
<td><strong>Who should be served by an evaluation?</strong></td>
</tr>
<tr>
<td><strong>What is the process of doing an evaluation?</strong></td>
</tr>
<tr>
<td><strong>What methods of enquiry should be used in evaluation?</strong></td>
</tr>
<tr>
<td><strong>Who should do evaluation?</strong></td>
</tr>
<tr>
<td><strong>By what standards should evaluation be judged?</strong></td>
</tr>
</tbody>
</table>
What is the evaluative context of problem-based learning?

Context is crucial in educational evaluation (Stufflebeam & Phi Delta Kappa National Study Committee on Evaluation 1971). Here, the term PBL refers to the overall ‘problem-first’ philosophy guiding a curriculum, and to the process for small-group PBL sessions as one element of that curriculum (Barrows 1986; Engel 1991; Schmidt 1993), both setting the context for evaluation:

In the problem-first philosophy, the learning sequence places the reason/trigger for needing to know something (its practical application) first, to stimulate the students’ learning about it. A teacher-centred philosophy would place being told what to learn first, followed by how to apply it.

In PBL sessions, the process involves students generating their own learning objectives from facilitated small-group-work on clinical scenarios in a safe, collaborative atmosphere of free enquiry that encourages efficient acquisition, handling and critical appraisal of knowledge, and self-evaluation and reflective practice (Maudsley 1999a).

All three major reviews of the outcomes of PBL vs. traditional approaches to undergraduate medical education (Albanese & Mitchell 1993; Berkson 1993; Vernon & Blake 1993) struggled, predictably, with how PBL and problem-based curricula were defined. They focused on effectiveness over the previous two decades:

- Albanese and Mitchell’s meta-analysis-type review accompanied curricular renewal in Iowa Medical School, identifying relevant questions by informal interviews with faculty and administration colleagues. They quantified ‘effect sizes’, e.g.: students’ basic science and clinical examination performances; student satisfaction, selection and retention; students’ study behaviours; and clinical ratings of graduates and undergraduates (reinforced by subsequent evidence, Richards et al. 1996). They found PBL to be more nurturing and enjoyable, with graduates performing at least as well as traditional counterparts and tending more towards family medicine. Flaws in the evidence-base precluded firm conclusions, but potential concerns involved: over-dependence on small-group-working, cost, and unsatisfactory basic science content coverage and related ‘cognitive scaffolding’.
- Vernon and Blake’s meta-analysis found consistently more positive programme evaluation, attitudes, attendance, mood and clinical performance amongst PBL students, without knowledge deficiencies.
- Berkson’s literature review concluded that the ‘products’ of both curricular types appeared indistinguishable, possibly because of students’ adaptability, similar learning materials or flawed study design. She considered that PBL curricula were unlikely to surpass traditional curricula in imparting ‘problem-solving’, imparting knowledge, enhancing motivation, improving self-directed learning, or providing happier experiences. Berkson forgave higher costs and failure to meet unrealistic expectations, however, because PBL is committed explicitly to the students and educational principles yet still manages to produce competent graduates.

These three reviews differed in purpose, study selection methods and ‘outcome’ indicators, and could only be cautiously optimistic about recommending wholesale conversion to problem-based curricula. Nevertheless, they all acknowledged philosophical advantages of PBL through valuing explicitly student and adult learning principles.

Given that students should reflect on and evaluate individual and group progress at the end of each PBL session (Maudsley 1999b), their PBL tutors have the potential to be reflective role-models. Indeed, maybe these educators can contribute to a more receptive educational context for ‘holistic’ programme evaluation.

What philosophy has guided programme evaluation of innovative undergraduate medical curricula? Examples

Referring to the early years of the McMaster undergraduate medical curriculum, Hamilton (1992 p. 1) warned against complacency in innovative curricula:

... one senior epidemiologist who should have known better said there was no need to evaluate because other curricula had failed and McMaster necessarily had to do better.
How should they be evaluated? The randomized controlled trial (RCT) is the quantitative ‘gold standard’ study underpinning evidence-based health care, minimizing bias through randomly allocating ‘exposure’. Nevertheless, RCTs may well be inappropriate, untenable and unethical in many educational settings, e.g. at whole curriculum-level. Harvard, for example, was able to compare students/graduates by RCT because it had twin-track traditional/innovative curricula for a short time (Block & Moore 1994). Indeed various schools have used the twin-track strategy (Mennin et al. 1996), sometimes to provide evaluation evidence that encourages the innovative track and main track subsequently to meld (Suwanwela et al. 1993).

Suitable definitions and outcome measures for comparing medical graduates of traditional and innovative curricula remain elusive in the few reports of consensus-seeking exercises, e.g. from Chapel Hill, North Carolina (2-day conference of eight experts in 1989) (Friedman et al. 1990) and from Australia (2-day workshop in 1991) (Pearson & Wallis 1992). Both reports highlighted the need to raise generalizable, important questions about the graduates, e.g. for lifelong learning. Friedman et al. (1990) suggested: Do they know the latest? Do they influence the practice behaviour of colleagues? These reports differed, however, on whether to advocate using routinely collected data (as a pragmatic response to the under-funding of evaluation) (Pearson & Wallis 1992) or collecting data fit for purpose (Friedman et al. 1990).

Friedman’s definition of innovative curricula excluded single-‘subject’ problem-based courses within otherwise traditional medical curricula (although two influential overviews of evaluation evidence included these to retain data, Vernon & Blake 1993; Albanese & Mitchell 1993). Friedman et al. (1990) were also focusing towards the curricular product when concluding that important evaluation studies:

- emphasize outcomes, i.e. the ‘distinctive product’;
- explore non-ephemeral outcomes;
- can be concluded within a decade;
- will interest decision-makers involved in changing to or continuing innovative curricula;
- do not duplicate previous studies unnecessarily; and
- cannot be performed routinely.

Friedman et al. (1990) hypothesized that there were five possible indicators on which students or alumni of innovative and traditional curricula were not expected to differ substantially, of which they recommended focusing on two: clinical problem-solving in common uncomplicated disease, and educational costs per student (excluding ‘start-up’). The assumption, however, that these should be similar appears debatable. Friedman et al. expected differences, however, on 26 indicators in six clusters: (i) psychosocial/interpersonal skills; (ii) continued learning; (iii) professional satisfaction; (iv) practice behaviour; (v) educational achievement/cognitive development; and (vi) institutional issues. They recommended focusing on the first three clusters.

The Chapel Hill work highlighted the difficulties in evaluating innovative undergraduate medical curricula: study design; inability to combat non-random allocation; contamination effects (e.g. of postgraduate education); and time trends (Friedman et al. 1990).

The Newcastle workshop considered programme evaluation in Australian medical schools, acknowledging the politics of evaluation (Pearson & Wallis 1992) and the scarcity of external funding even in innovative schools (Smith 1992a). Nevertheless, Hamilton (1992 p. 4) noted the:

...considerable value in an informed, shrewd and dispassionate descriptive appraisal of medical education.

The ‘dispassionate’ label is consistent with what Stronach referred to as a ‘hands-off’ approach to evaluation (rather than ‘hands-on’ or ‘hands-up’ approaches) (Stronach & Torrance 1995), i.e. an objectivity shared by health care evaluation (Donabedian 1966): How to evaluate defied consensus. Woodward (1992 p. 7) warned, however, against invariably pursuing outcome evaluation, as cost might outweigh the information obtained, noting that:

If no actions will be taken (even the decision to maintain the status quo) by the answers we receive, the question is not worth asking.

Smith summarized that outcome evaluation should highlight aspects emphasized by a curriculum, i.e.
expected differences (Smith 1992a, 1992b). Furthermore, if curricula are organized around vertical themes, these should guide student assessment and programme evaluation.

Predictably, Gerrity and Mahaffy’s (1993) review of the state of evaluation in eight North American medical schools undergoing reform found that most had delayed their programme evaluation; all had used mixed approaches. They concluded that programme evaluation:
- can stabilize a turbulent political environment;
- becomes increasingly important in its formative mode as the curricular change becomes more comprehensive and complex;
- has much to gain from incorporating qualitative approaches;
- can be overwhelmed with excessive data collection;
- must close the feedback loop; and
- can stimulate curricular change when centralized.

On the last point, such centralization of programme evaluation is not straightforward (Henderson et al. 1993). Block & Moore (1994 p. 114) blamed the ‘dearth of reliable evaluations’ for perpetuating traditional dogma, a warning sounded beyond medicine as well. From veterinary studies in Melbourne, for example, McNaught et al. (1999 p. 248) reported how ‘inadequate evaluation, isolated innovation, and fragmented institutional policy’ prevented problem-based learning from being used beyond one initial short-course in veterinary bacteriology and mycology. Only a subsequent change of university management and growing international consensus about the value of PBL allowed a centralized programme-wide policy to promote a more widespread implementation.

Considered explorations of the philosophy underpinning undergraduate medical programme evaluation are unusual, and could benefit from debates about general educational and health care evaluation.

**What can educational and health care evaluation frameworks offer?**

In terms of its function and focus, Brookfield (1986) summarized five main approaches to educational evaluation for adult learning, which differ in their emphasis (Box 2). Firstly, systematic evaluation against the planned educational objectives (as promoted by Tyler et al. 1989) provides a structured but narrow approach. Secondly, Scriven’s (1967) goal-free evaluation takes account of unplanned as well as planned outcomes, thus valuing more than the educators’ initial intentions. Thirdly, Stufflebeam et al.’s (1971) ‘CIPP’ evaluation of Context, Input, Process, Product, widened the scope beyond outcomes. Fourthly, Kirkpatrick’s (1967) hierarchy of

**Box 2 Five approaches to educational evaluation according to Brookfield. Source: summarized from Brookfield (1986)**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyler’s (1989) evaluation, i.e. against the planned educational objectives:</td>
<td>. . . attracts criticism about: being school-orientated, disregarding unplanned outcomes, being authoritarian, over-emphasizing programme objectives, being inflexible, and ignoring the diversity of students’ experiences, levels, etc.</td>
</tr>
<tr>
<td>Scriven’s (1967) goal-free evaluation:</td>
<td>. . . responds to the above pitfalls by aiming to be more student-centred, and acknowledging unintended outcomes important to students.</td>
</tr>
<tr>
<td>Stufflebeam et al. (1971) ‘CIPP’ evaluation focused on Context, Input, Process, Product:</td>
<td>. . . without working to the explicit aims of the programme, an evaluator reports on its main achievements and effects, which can then be compared with the planned educational objectives and set in the context of hidden agendas and assumptions.</td>
</tr>
<tr>
<td>Kirkpatrick’s (1967) hierarchy of evaluation:</td>
<td>. . . focuses on pre-determined objectives, but does value effects beyond the base institution when it places the evidence at four levels (ascending): reaction of programme participants, learner satisfaction; knowledge and skills achieved; transfer of learning from classroom to practice; the community impact (beyond the institution).</td>
</tr>
<tr>
<td>Guba’s (1978) naturalistic evaluation:</td>
<td>. . . focuses on starting with the evaluation questions and criteria of the participants, reporting back to them, acknowledging political influences and context, and improving their experience as a result, using qualitative research approaches.</td>
</tr>
</tbody>
</table>
evaluation highlighted outcomes beyond the institution by placing them at the top of four levels of evidence (learners’ satisfaction, their learning, its transfer to practice, the community impact). Evaluating at the lowest level only, i.e. student satisfaction, is unhelpful:

... learners expressing dissatisfaction may do so because the program has caused them to engage in a difficult and possibly painful examination of the unquestioned assumptions by which they are living or working.

(Brookfield 1986 p. 271)

Guba’s (1978) naturalistic evaluation starts with the concerns of students and reports findings to them, trying to improve their experience. Collectively, these five approaches provide guidance for educational evaluation: to refer to the planned objectives; the unplanned outcomes; the input/context/process; the impact in the workplace and wider community; and the participants’ perspective, as appropriate.

Turning to health care evaluation, doctors often ask ‘Yes... but will it make a difference... [to the individual patient]?’ about proposed changes in medical practice. Their public health medical colleagues, as the evaluators amongst medical specialties (Thomson & Bhopal 1993), ask the similar ‘Yes... but so what [for the population’s health]?’ question about various health interventions, while taking a population overview of health and health care. When the 1990s NHS reforms promoted the measurement of health care quality (Hopkins 1990; Thomson & Bhopal 1993; Secretary of State for Health 1997), the evaluative framework of Donabedian (1966) and the quality dimensions of Maxwell (1984) informed the ensuing debate on ‘How?’.

Donabedian’s framework for health care evaluation assessed:
- structure (‘settings and instrumentalities’, p. 170);
- process (care practices known to be ‘good’); and
- outcome (results achieved).

This pre-dated Stufflebeam et al. CIPP approach in educational evaluation, which likewise considered issues other than just outcomes. Although Donabedian focused originally on the physician-patient level, explicitly excluding the population level, public health workers tend to find this framework useful at the population level. When Donabedian (1966 p. 189) likened the search for simple tools to evaluate complex health care to ‘pursuing a will-o’-the-wisp’, he might as well have been referring to educational evaluation. Either type of evaluation is context-dependent, multidimensional, and not necessarily best judged by outcomes alone:

... the definition of quality may be almost anything anyone wishes it to be, although it is, ordinarily, a reflection of values and goals current in the medical care system and in the larger society of which it is a part.

(Donabedian 1966 p. 167)

Nevertheless, Donabedian’s ‘ultimate validators of... [the] quality of medical care’ (p. 169), used appropriately, were outcome indicators of health (i.e. effectiveness) and satisfaction.

In contrast, Maxwell (1984 p. 1471) asserted that simple health care quality assessment should ‘not be beyond human wit’. Indeed, Maxwell’s three ‘A’s and three ‘E’s of quality could usefully inform undergraduate medical educational evaluation:
- acceptability (social);
- accessibility;
- appropriateness (relevance to population need);
- effectiveness (for individual ‘consumers’);
- efficiency; and
- equity.

Like most medical educational evaluation, which tends to under-emphasize acceptability, accessibility, appropriateness and equity of educational provision, Maxwell and Donabedian both highlighted outcomes. In health care evaluation, Crombie and Davies reviewed the neglect of process-measurement at the expense of outcome-measurement (Davies & Crombie 1997; Crombie & Davies 1998). They noted that process-indicators have higher sensitivity and are easier to measure, easier to interpret (e.g. without having to compare; without having to adjust for case-mix), and often point directly to specific deficiencies in care. The art-science tension that bedevils health care evaluation (Grémy 1999) is also evident in educational evaluation. Health care evaluation therefore can offer alternative ways of conceptualizing educational evaluation; they share the difficulty of making sense out of ‘messy’ situations and systems. The medical literature tends not to highlight such mutually beneficial connections between these literatures, but does raise further practical issues for consideration.
What can be learned from examples of published accounts of problem-based undergraduate medical programme evaluation?

There are numerous ways in which programme evaluation could be tackled. Stake classified evaluation designs along eight dimensions (Stake 1976):

- formative–summative;
- formal–informal;
- case-particular–generalization;
- product–process;
- descriptive–judgemental;
- pre-ordinate–responsive;
- holistic–analytical; and
- internal–external.

As inferred previously, the educational climate for undergraduate medical curricula encourages formal, generalizable, product-orientated, judgemental, pre-ordinate study that is usually conducted internally and analytically, and while it can be summative (i.e. it 'counts'), the focus here is on the formative (for feedback) purpose. Wilkes & Bligh (1999) classified approaches to evaluating undergraduate medical education as:

- student-orientated (mostly student performance);
- programme-orientated (mostly judging how elements/processes contribute to overall outcomes);
- institution-orientated (mostly comparative analysis by external bodies); and
- stakeholder-orientated (mostly the claims and concerns of students, staff, patients and health service).

Here, the focus is on the programme-orientated approach, of which reported examples have been less obvious from traditional (El-Hazmi & Haque 1985) compared with innovative (Pinto Pereira et al. 1993) curricula.

As educational evaluation should be context-specific, there will be no one ‘best’ approach, as revealed by the literature. Ross & Fineberg (1998) examined the impact of 10 North American undergraduate medical curricula that were in place by 1990 and ‘known as’ innovative. They reported from group interviews that, irrespective of whether the schools were smaller and education-service-orientated or larger and education-research-service-orientated, students shared an appreciation of certain curricular characteristics. These comprised the promotion of: professional and personal individuation; connection to students, faculties, patients, and communities via collaborative work; and diversity.

In their review of eight North American medical schools, Gerrity & Mahaffy (1998) found an eclectic evaluation culture, with the focus on internal, formative and quantitative evaluation. Predominantly qualitative approaches (e.g. participant observation, Haas & Shaffir 1982) are unusual.

Sherbrooke and Hawaii provided rare published accounts of successful ‘one step’ conversion from traditional to problem-based format. These medical schools considered such transformation to be ‘... neither unrealistic nor unattainable’ (Des Marchais 1991 p. 237) and critically reliant on tutor training (Anderson 1991; Des Marchais 1991, 1993; Grand’Maison & Des Marchais 1991). Personal experience of the PBL tutorial appeared pivotal in staff accepting the transformation. Reports from Sherbrooke highlighted the relationship between evaluation and change management (i.e. focusing on implementation), the importance of describing evaluation context, and the way in which learning theory could inform the design of evaluative instruments (Des Marchais et al. 1989; Des Marchais 1991, 1993). Including staff among the ‘objects’ of evaluation, Des Marchais (1991) reported a high participation rate in tutor training, increased interest in medical education, modification of educational philosophy, and successful acquisition of specific skills by staff. Others have reinforced this crucial emphasis on staff development (Evans & Taylor 1996). Two other examples of evaluative studies warned against staff complacency:

- A questionnaire survey compared the attitudes of teaching staff at a traditional (Cardiff) and an innovative (Newcastle, Australia) medical school to teaching and teacher training (Finucane et al. 1995). Both found that educators over-estimated their personal ability, perceived a lack of rewards and were ambivalent towards formal training.
- An interview study of the foundation staff of a problem-based curriculum indicated decreasing enthusiasm for PBL after a few years, and the tendency to revert to classical didactic teaching.
after early pioneers had left (Abdulrazzaq & Qayed 1991).

Sherbrooke reported a three-stage formative programme evaluation, over a 6-year period, using external evaluators and a combined 'expert' judgement and 'connoisseur' model of programme evaluation (Des Marchais & Bordage 1998). The evaluation focused initially on increasing the profile of the reform, then on the first cohort of graduates, and finally on promoting curriculum self-renewal. Brandon et al. (1992, 1993) reported that involvement of students in setting the evaluation agenda at Hawaii led to widening the focus beyond merely outcomes to other factors in the learning environment such as communication and other affective elements.

Truly reflective evaluation is difficult in the mainly positivistic, quantitative environment of a medical school. Kember & Gow (1992) recommended better use of the 'plan–act–observe–reflect' cycle of action research for curricular improvement in higher education generally, but this is rarely applied to undergraduate medical curricula. Prideaux (1993) reported a failed attempt to use action research as an agent of change in the surgery part of the 'self-directed learning' curriculum at Flinders (South Australia). Al-Haddad & Jayawickramarajah (1991) reported using 'action research' when they surveyed hospital supervisors of new graduates of the Bahrain problem-based curriculum (with a questionnaire modified from Newcastle, Australia), but their definition was rather loose.

Irrespective of approaches (qualitative and/or quantitative), Nevo (1986) contended that all evaluation processes should include three activities: focusing on the evaluation ‘problem’; collecting and analysing empirical data; and communicating findings to the evaluation audience. Even for problem-based undergraduate medical curricula, however, it is sometimes difficult to reconstruct meaningfully the specific evaluation problem and philosophy from published reports. Indeed, the ‘true picture’ probably loses something in translating between intended audience and academic publication. Acknowledging such constraints, what can be learned from examples of published accounts in this field?

Newcastle (Australia) reported its explicit commitment to programme evaluation (Feletti & Fisher 1979; Rolfe & Pearson 1997) with its first medical student intake of 1978 (Engel et al. 1980), recognizing evaluation to be integral to a faculty founded on innovation and continuing improvement (Feletti & Fisher 1979). A Programme Evaluation Committee and a faculty primarily structured around functional not disciplinary groups (e.g. student assessment, admissions, programme evaluation, etc.) underpinned this. Early on, Newcastle used Stufflebeam et al. ‘CIPP’ framework (Feletti & Fisher 1979) to focus limited resources on the planning and implementation dimensions of the Context, Implementation, Planning and Product (educational outcomes) framework. Feletti & Fisher (1979) identified the difficulties as:

- deciding on the aspects to evaluate;
- the need for rapid information flow;
- the degree of co-operation needed from staff and students;
- the adequacy of information returned; and
- sampling over time.

They warned against reacting prematurely to staff/student perceptions:

> It is only too easy for the consequences of rapid and spontaneous minor changes, made in response to formative evaluation, to be overlooked in relation to their influence on the more substantive and long-term programme objectives.

(Feletti & Fisher 1979 p. 318)

They also warned of the dangers of initial planning excitement evaporating as implementation becomes more mundane (they could easily have widened this to the fate of evaluative intentions as well!). A decade later, evaluation became more product-focused when a critical mass of Newcastle graduates had accrued (Pearson & Wallis 1992).

The further issues raised by evaluating problem-based curricula relate to ensuring that sufficient effort is invested in timely evaluation that will maintain the momentum of curriculum renewal. Evaluators should mix their approaches, appropriate to special curricular features, and acknowledge specific purpose and likely caveats.

Where to from here?

As part of the curricular planning cycle, evaluation should stimulate the further development of
problem-based undergraduate medical programmes. The approach needs to be reflective, iterative and progressive, highlighting the special features of that curriculum. In comparing how eight medical schools approached evaluation of their curricular reforms, Gerrity & Mahaffy (1998 p. S56) considered that ‘Many schools found that they were building the roads as they traveled them’.

As problem-based education provides a more holistic conception of education (Margetson 2000), then it should be judged in a more holistic way. Despite shared philosophy, problem-based undergraduate curricula differ considerably from each other, hence the necessarily eclectic approach to programme evaluation. Even where PBL guides the overall programme design, the specific version of PBL can focus explicitly on promoting problem-solving and/or on promoting a useful and accessible knowledge-base. Papa & Harasym (1999), for example, have described North American curriculum design as developing through five models: apprenticeship (1765–); discipline-based (1871–); organ-system-based (1951–); problem-based (1971–); giving way to the clinical-presentation-based model (1991–) at Calgary to de-emphasize generic problem-solving and emphasize structured-knowledge formats.

Programme evaluation cannot conform to a strict ‘recipe’-type approach, given differences in: the question being asked about a particular problem (the function of the evaluation), who is asking, and local context/needs. In reviewing theory and evidence about the adult learner, Knowles (1990) noted that educational evaluation should be primarily about improving teaching and learning, not merely justifying current practice. Ideally, evaluators should:

- Articulate the reason for the evaluation and how the design could be characterized, for example, in terms of Nevo’s (1986) and Stake’s (1976) dimensions, and report this explicitly along with ‘the findings’.
- Be realistic, design the evaluation consistent with the local context and special features, and not ignore the unplanned outcomes (and underlying process) of innovation in problem-based curricula: ‘Outcome... goes beyond the normal interpretation of the product of the educational process, namely, the students as graduates. It also includes the academic staff... [measuring its growth] as an instructional group or faculty resource. Outcome can also include the educational materials that have been developed...’ (Feletti & Fisher 1979 p. 317).
- Plan a portfolio of activities, running throughout the curriculum, which reflect design features such as thematic orientation in truly integrated programmes.
- Make appropriate use of qualitative and quantitative approaches, consistent with the question being asked.
- Draw upon health care evaluation frameworks (Donabedian 1966; Maxwell 1984), as appropriate, where these offer alternative ideas, especially in considering acceptability, accessibility, appropriateness and equity of provision, as well as effectiveness and efficiency.

Using ‘gold standard’ RCT evidence to establish ‘causality’ between specific curricular characteristics and outcomes is probably unrealistic. ‘Routine’ data collection is only justifiable for explicit aims. Even with databases set up specifically to provide evaluation data from the outset, it might prove difficult and time-consuming to rationalize why certain items were ever collected at all (Murphy 1992).

British undergraduate medical programme evaluation is starting to gain greater priority. National imperatives are: re-orientating medical education; updating health professional regulation in the interests of protecting the patient/public and improving service delivery; and quality-assuring higher education. The Dearing recommendations about accrediting university educators and developing learning outcome specifications for all educational programmes could potentially enhance programme evaluation.

Problem-based medical schools, with staff particularly active in educational discourse, could contribute substantially to advances in programme evaluation (Fisher 1991). Nevertheless, unrealistic expectations to design and fulfill outcome indicators of ‘true’ educational quality, without challenging the ‘legitimacy’ of traditional curricula and traditional evaluation-measures, are unhelpful. It is, for example, debatable whether an evaluation question like ‘Are our graduates better physicians?’ (Smith & Dollase 1999) can ever be answered satisfactorily; maybe it should be
rephrased as ‘What makes our graduates competent doctors?’.

Halperin et al. (1995) highlighted the need for North American medical schools to move beyond insularity and share their expertise with, and learn from, other university academic disciplines. It is notable, however, that in citing the educational strengths that each could learn from the other, Halperin et al. did not highlight programme evaluation. Nevertheless, they did highlight that the combination of health care delivery and undergraduate medical education made fiscal and regulatory accountability a strength worth sharing with the remainder of the university sector. Maybe public health academics could usefully contribute to educational evaluation and relate achievements to potential impact on population health and health services. ‘Healthier links’ are needed between the health care evaluation and educational evaluation literature. There is almost a literature compartment syndrome to combat.

Being clear about the meaning and expectations of programme evaluation is a useful starting point for future work. Effort expended chasing elusive outcomes needs complementing with meaningful interpretation of process. Donabedian’s (1966) advice on health care evaluation, i.e. that understanding health care process is as important as judging its quality, is transferable to educational evaluation. Will health care evaluation and educational evaluation, medical schools (and the remainder of the university sector) and the NHS learn from each other? Recent high-profile investigations into clinical, professional and managerial practice at the academic–NHS interface (Smith 2001) indicated the need for more collaborative and shared learning. Clinical governance and educational governance (Morrison & Buckley 1999) for undergraduate medical education are related issues. It remains to be seen whether the medical educational changes that the GMC’s (1993) Tomorrow’s Doctors triggered in this country will stimulate important advances in educational evaluation.

Acknowledgement

Thank you to Mr Ron Stewart for guidance and thought-provoking discussions on the Master of Education module that originally stimulated and organized my ideas about this topic.

References


